Date: 15.10.24

Statistics Worksheet – 1

Answers

1. a)

2. a)

3. d)

4. d)

5. c)

6. b)

7. b)

8. a)

9. c)

10. The term Normal Distribution refers to a probability distribution of continuous random variable. It is also known as Gaussian Distribution. It is characterized by its mean equal to 0 and standard deviation equal to 1. The normal distribution curve is a bell shaped curve. The distribution is symmetric and approximately 68 percent of the data collected will fall within +/- one standard deviation of the mean. Two examples of normal distribution are the amount of rainfall in inches in a year for a city and the weight of a newborn baby.

11. Missing data are represented as NaN (Not a Number) in many programming languages including Python. The two main approaches to handle missing data are as follows:

a) Deletion – this involves removing rows or columns with missing values.

b) Imputation – this replaces the missing values with estimates. The various imputation techniques are as follows:

i) Mean/Median/Mode imputation – replace missing value with average, middle or most frequent value of the corresponding column.

ii) K-Nearest Neighbors (KNN imputation) – this method finds the closest data points (neighbors) based on available features and uses their values to estimate the missing value. KNN is useful when there is a lot of data and the missing values are scattered.

iii) Model-based imputation – this involves creating a statistical model to predict the missing values based on the other features in the data. This can be a powerful technique; but it requires more expertise and can be computationally expensive.

12. A/B testing is a type of experiment in which the web traffic or user base is split into two groups, and which shows two different versions of a web page, app, email and so on, with the goal of comparing the results to find the more successful version. With an A/B test, one element is changed between the original and the test version to see if this modification has any impact on user behavior or conversion rates. From a data scientist's perspective, A/B testing is a form of statistical hypothesis testing or a significance test.

13. The mean imputation of missing data is a quick and easy approach, but it can introduce bias if the missing data is not randomly distributed.

14. Linear Regression uses one independent variable to explain or predict the outcome of the dependent variable Y. The linear regression equation is Y = a + bX + e

where,

Y is the dependent variable,

X is the independent variable,

a is the intercept,

b is the slope, and

e is the regression residual error

15. The two main branches of statistics are descriptive statistics and inferential statistics.